DOCKET NO.: 15675P538

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re the Application of:

ENRICO MAIM

Application No.:

Filed:

For: **methods and systems for searching and associating information resources such as web pages**

Art Group:

Examiner:

Commissioner for Patents
P.O, Box 1450
Alexandria, VA 22313-1450

REQUEST FOR PRIORITY

Sir:

Applicant respectfully requests a convention priority for the above-captioned application, namely:

COUNTRY	APPLICATION NUMBER	DATE OF FILING
France	0200341	11 January 2002
France	0205751	7 May 2002

☐ A certified copy of the document is being submitted herewith.

Respectfully submitted,

Blakely, Sokoloff, Taylor & Zafman LLP

Dated: 7/12/04

Eric S. Hyman, Reg. No. 30,139

12400 Wilshire Boulevard, 7th Floor
Los Angeles, CA 90025
Telephone: (310) 207-3800

INPI
INSTITUT NATIONAL DE LA PROPRIETE INDUSTRIELLE

BREVET D'INVENTION

CERTIFICAT D'UTILITÉ - CERTIFICAT D'ADDITION

COPIE OFFICIELLE

Le Directeur général de l'Institut national de la propriété industrielle certifie que le document ci-annexé est la copie certifiée conforme d'une demande de titre de propriété industrielle déposée à l'Institut.

Fait à Paris, le 20 JAN. 2003

Pour le Directeur général de l'Institut
national de la propriété industrielle
Le Chef du Département des brevets

Martine PLANCHE

DOCUMENT DE PRIORITÉ

PRÉSENTÉ OU TRANSMIS CONFORMÉMENT À LA RÈGLE 17.1.a) OU b)

CERTIFIED COPY OF PRIORITY DOCUMENT

INSTITUT NATIONAL DE LA PROPRIETE INDUSTRIELLE
SIEGE
26 bis, rue de Saint Petersbourg
75800 PARIS cedex 08
Téléphone : 33 (1) 53 04 53 04
Télécopie : 33 (1) 42 93 59 30
www.inpi.fr

INPI
INSTITUT NATIONAL DE LA PROPRIÉTÉ INDUSTRIELLE

26 bis, rue de Saint Pétersbourg
75800 Paris Cedex 08
Téléphone : 01 53 04 53 04 Télécopie : 01 42 94 86 54

BREVET D'INVENTION
CERTIFICAT D'UTILITÉ

Code de la propriété intellectuelle - Livre VI

cerfa N° 11354*01

REQUÊTE EN DÉLIVRANCE 1/2

Important ! Remplir impérativement la 2ème page.

Cet imprimé est à remplir lisiblement à l'encre noire DB 540 W /190600

Réservé à l'INPI

REMISE DES PIÈCES DATE 11 JAN 2002
LIEU 75 INPI PARIS B
N° D'ENREGISTREMENT NATIONAL ATTRIBUÉ PAR L'INPI 0200341
DATE DE DÉPÔT ATTRIBUÉE PAR L'INPI 11 JAN. 2002

Vos références pour ce dossier (facultatif) NeoDistill

1 NOM ET ADRESSE DU DEMANDEUR OU DU MANDATAIRE À QUI LA CORRESPONDANCE DOIT ÊTRE ADRESSÉE

Enrico MAIM
17 rue Biscornet
75012 PARIS

Confirmation d'un dépôt par télécopie ☐ N° attribué par l'INPI à la télécopie

2 NATURE DE LA DEMANDE	Cochez l'une des 4 cases suivantes
Demande de brevet	☒
Demande de certificat d'utilité	☐
Demande divisionnaire	☐
Demande de brevet initiale	N° Date ___/___/___
ou demande de certificat d'utilité initiale	N° Date ___/___/___
Transformation d'une demande de brevet européen *Demande de brevet initiale*	☐ N° Date ___/___/___

3 TITRE DE L'INVENTION (200 caractères ou espaces maximum)

Procédés et systèmes pour identifier des contenus pertinents vis-à-vis de contenus donnés, et procédés et systèmes de navigation sur la Toile associés.

4 DÉCLARATION DE PRIORITÉ OU REQUÊTE DU BÉNÉFICE DE LA DATE DE DÉPÔT D'UNE DEMANDE ANTÉRIEURE FRANÇAISE	
	Pays ou organisation Date ___/___/___ N°
	Pays ou organisation Date ___/___/___ N°
	Pays ou organisation Date ___/___/___ N°
	☐ S'il y a d'autres priorités, cochez la case et utilisez l'imprimé «Suite»

5 DEMANDEUR		☐ S'il y a d'autres demandeurs, cochez la case et utilisez l'imprimé «Suite»
Nom ou dénomination sociale		MAIM
Prénoms		Enrico
Forme juridique		
N° SIREN		. . . . .
Code APE-NAF		. . . .
Adresse	Rue	17, rue Biscornet
	Code postal et ville	75012 PARIS
Pays		France
Nationalité		Italie
N° de téléphone *(facultatif)*		06 80 30 83 00
N° de télécopie *(facultatif)*		
Adresse électronique *(facultatif)*		

INPI
INSTITUT NATIONAL DE LA PROPRIÉTÉ INDUSTRIELLE

BREVET D'INVENTION
CERTIFICAT D'UTILITÉ

REQUÊTE EN DÉLIVRANCE 2/2

Réservé à l'INPI

REMISE DES PIÈCES 18 JAN 2002
DATE
LIEU 75 INPI PARIS B
N° D'ENREGISTREMENT NATIONAL ATTRIBUÉ PAR L'INPI 0200341

08 540 W /190800

Vos références pour ce dossier : *(facultatif)*	Neodistill
6 MANDATAIRE	
Nom	
Prénom	
Cabinet ou Société	
N °de pouvoir permanent et/ou de lien contractuel	
Adresse — Rue	
Adresse — Code postal et ville	
N° de téléphone *(facultatif)*	
N° de télécopie *(facultatif)*	
Adresse électronique *(facultatif)*	
7 INVENTEUR (S)	
Les inventeurs sont les demandeurs	[x] Oui [ ] Non **Dans ce cas fournir une désignation d'inventeur(s) séparée**
8 RAPPORT DE RECHERCHE	**Uniquement pour une demande de brevet (y compris division et transformation)**
Établissement immédiat ou établissement différé	[x] [ ]
Paiement échelonné de la redevance	**Paiement en deux versements, uniquement pour les personnes physiques** [x] Oui [ ] Non
9 RÉDUCTION DU TAUX DES REDEVANCES	**Uniquement pour les personnes physiques** [ ] Requise pour la première fois pour cette invention *(joindre un avis de non-imposition)* [x] Requise antérieurement à ce dépôt *(joindre une copie de la décision d'admission pour cette invention ou indiquer sa référence)* :
Si vous avez utilisé l'imprimé «Suite», indiquez le nombre de pages jointes	

10 SIGNATURE DU DEMANDEUR OU DU MANDATAIRE (Nom et qualité du signataire)

Enrico MAIM

VISA DE LA PRÉFECTURE OU DE L'INPI

M. MARTIN

La loi n°78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés s'applique aux réponses faites à ce formulaire.
Elle garantit un droit d'accès et de rectification pour les données vous concernant auprès de l'INPI.

La présente invention concerne un système et procédé associé destinés à assurer les fonctions suivantes :

- recevoir en entrée une *requête de recherche* composée d'un ensemble d'identificateurs URI[1] (ou URL[2]), ces identificateurs permettant d'accéder à des ressources d'information telles que des pages (pages Web) ou des parties de page ;
- fournir à un utilisateur les URI (ou directement les pages) qui sont censés être les plus pertinents par rapport à ladite requête ;
- optionnellement, permettre à l'utilisateur d'ajouter ensuite à sa requête les URI qu'il a effectivement appréciés et le système peut ainsi affiner sa réponse itérativement.

La requête peut par exemple être constituée des liens favoris de l'utilisateur, le but du système étant de surveiller le Web autour de ces liens et de notifier l'utilisateur quand de nouvelles pages intéressantes y apparaissent, soit en technologie « Push » à l'initiative d'un serveur, soit en technologie « Pull » à l'initiative de l'utilisateur.

Le système est d'abord basé sur une analyse des liens hypertextes des pages Web. Grâce à cela, il peut fournir de bons résultats même s'il n'y a qu'un seul utilisateur.

Néanmoins, le système peut aussi exploiter les requêtes des utilisateurs. Notamment le système peut exploiter les regroupements d'URI effectués par chaque utilisateur dans ses requêtes (par exemple dans ses répertoires de liens favoris).

Le système peut aussi détecter la proximité de centres d'intérêt d'utilisateurs, ou encore leurs complémentarités (i.e. pour ou contre le projet présenté, acheteur ou vendeur du produit présenté, etc), grâce par exemple à des étiquettes d'attributs associées aux contenus informationnels présentés. L'utilisateur fournit ainsi au système sa « position » par rapport au contenu présenté, et le système peut ainsi mettre en relation non pas seulement les utilisateurs ayant des centres d'intérêt proches mais aussi ceux qui ont une position complémentaire par rapport à un contenu.

De manière plus générale, le procédé peut être adapté pour être appliqué à la recherche de n'importe quelle ressource d'information (pas seulement des pages Web), notamment sur un réseau. Au lieu d'exploiter les liens hypertextes et les requêtes comme mentionnés ci-dessus, le système peut être basé sur une analyse des traces des copier-coller (ou couper-coller) de fragments d'information effectués par les utilisateurs (dans le cadre des créations et manipulations de ressource d'information), pour suggérer automatiquement d'autres fragments qui sont susceptibles d'enrichir ces ressources. Ces traces peuvent en effet être assimilées à des liens. Par exemple, quand on copie une partie d'une page Web dans un document, le système est capable d'en déduire et de mémoriser l'existence d'un lien entre la page Web et le document, et les mêmes mécanismes décrits ici peuvent alors être appliqués.

L'invention propose tout d'abord un procédé pour déterminer des pages additionnelles pertinentes par rapport à un ensemble donné de pages de départ, caractérisé en ce qu'il comprend les étapes suivantes :

a) identifier un ensemble de pages citantes constituées par toutes les pages ayant un lien vers au moins l'une des pages de départ,

---

[1] (URI : Universal Resource Identifier) Dans la suite on utilise les termes « lien » (« lien hypertexte ») ou URI indifféremment (souvent de manière interchangeable) pour parler de l'identificateur d'une ressource (forme interne) ou de sa présentation dans la page affichée (forme externe).

[2] Uniform Resource Locator

b) former un ensemble de pages candidates constitué par l'ensemble des pages citées par les pages citantes,

c) pour chaque page candidate, calculer un score de pertinence de page candidate entre ladite page candidate et l'ensemble de pages de départ sur la base de l'existence de liens situés dans les pages citantes et dirigés vers la page candidate et vers les pages de départ, et sur la base également de scores de pertinence de pages citantes affectés à chacune des pages citantes,

d) pour chaque page citante, recalculer un score de pertinence de page citante sur la base de l'existence, dans la page citante en question, de liens vers les pages candidates et sur la base également des scores de pertinence de page candidate attribuées aux pages candidates à l'étape c),

e) répéter le cas échéant l'étape c) et le cas échéant une ou plusieurs fois l'étape d) puis l'étape c)

f) déterminer lesdites pages additionnelles pertinentes comme étant les pages candidates qui présentent les meilleurs scores de pertinence de page candidate.

Avantageusement, le calcul de score de pertinence effectué à l'étape c) comprend le calcul d'une pluralité de sommes de scores de pertinence de pages citantes, chaque somme comprenant uniquement les scores de pertinences des pages citantes comprenant un lien vers une page donnée constituée par la page candidate ou une page de départ.

De préférence, le procédé comprend également le calcul d'au moins une somme de scores de pertinence de pages citantes, chaque somme comprenant uniquement les scores de pertinences des pages citantes comprenant un lien vers l'une parmi un ensemble d'au moins deux pages données, cet ensemble comprenant la page candidate et au moins une page de départ.

L'invention propose également un procédé pour déterminer des pages additionnelles pertinentes par rapport à un ensemble donné de pages de départ, caractérisé en ce qu'il comprend les étapes suivantes

a) identifier un ensemble de pages citées constituées par toutes les pages ayant un lien depuis au moins l'une des pages de départ,

b) former un ensemble de pages candidates constitué par l'ensemble des pages citant les pages citées,

c) pour chaque page candidate, calculer un score de pertinence de page candidate entre ladite page candidate et l'ensemble de pages de départ sur la base de l'existence de liens situés dans la page candidate et dans les pages de départ et dirigés vers les pages citées, et sur la base également de scores de pertinence de pages citées affectés à chacune des pages citées,

d) pour chaque page citée, recalculer un score de pertinence de page citée sur la base de l'existence, dans la page citée en question, de liens depuis les pages candidates et sur la base également des scores de pertinence de page candidate attribuées aux pages candidates à l'étape c),

e) répéter le cas échéant l'étape c) et le cas échéant une ou plusieurs fois l'étape d) puis l'étape c)

f) déterminer lesdites pages additionnelles pertinentes comme étant les pages candidates qui présentent les meilleurs scores de pertinence de page candidate.

Selon un autre aspect, l'invention propose un système de navigation parmi des ressources d'information, chaque ressource comprenant au moins un lien activable dans un premier mode par un dispositif d'entrée pour provoquer l'accès à une autre ressource d'informations désignée par un identificateur de ressource associé à ce lien, caractérisé en ce qu'au moins certaines ressources comprennent au moins un lien activable dans un second mode à l'aide d'un dispositif

d'entrée pour envoyer à un moteur de recherche de nouvelles ressources d'informations une requête de recherche contenant l'identificateur de ressource associé au lien en question.

Avantageusement, le dispositif d'entrée est apte à activer le lien simultanément dans les premier et second modes.

De façon préférée, l'activation du lien dans le second mode est apte à provoquer l'affichage d'une requête pré-existante, à laquelle l'identificateur de ressource associé au lien en question est susceptible d'être ajouté.

Dans ce cas, le système est préférentiellement apte à afficher, en plus de la requête pré-existante, la ressource d'informations désignée par ledit identificateur de ressource.

Selon un autre aspect, l'invention propose un système de recherche de nouvelles ressources d'information à partir de ressources d'informations existantes, caractérisé en ce qu'il comprend un moteur de recherche basé sur l'analyse de liens entre différentes ressources et acceptant en entrée une requête comprenant une série d'identificateurs de ressources, caractérisé en ce qu'il comprend un moyen de sélection d'identificateurs apte à mémoriser un ensemble d'identificateurs (URI) de ressources sélectionnés les uns après les autres par un utilisateur, et un moyen générateur de requête activable par l'utilisateur pour engendrer une requête contenant l'ensemble des identificateurs précédemment sélectionnés à destination du moteur de recherche.

Avantageusement, le moyen de sélection est apte à mémoriser les identificateurs sélectionnés de manière rémanente, de telle sorte que le moyen de sélection puisse être mis en œuvre de façon espacée dans le temps en vue de la génération d'une même requête.

L'invention propose par ailleurs un procédé de recherche de nouvelles ressources d'information à partir de ressources d'informations existantes, caractérisé en ce qu'il comprend la mise en œuvre d'un moteur de recherche basé sur l'analyse de liens entre différentes ressources et acceptant en entrée une requête comprenant une série d'identificateurs de ressources et en ce qu'il comprend les étapes suivantes :

- sélection d'identificateurs (URI) de ressources les uns après les autres par un utilisateur ;
- génération d'une requête contenant l'ensemble des identificateurs précédemment sélectionnés à destination du moteur de recherche.

Enfin l'invention propose un procédé de recherche de nouvelles ressources d'information à partir de ressources d'informations existantes, caractérisé en ce qu'il comprend la mise en œuvre d'un moteur de recherche basé sur l'analyse de liens entre différentes ressources et acceptant en entrée une requête comprenant une série d'identificateurs de ressources et en ce qu'il comprend les étapes suivantes :

- génération d'une requête contenant un ensemble d'identificateurs de ressources précédemment mémorisés dans un même groupe d'identificateurs de ressources propre à un utilisateur, à destination du moteur de recherche,
- génération d'une signalisation à l'attention de l'utilisateur lorsqu'au moins un nouvel identificateur de ressource appartenant au groupe en question a été trouvé par le moteur.

De préférence, chaque groupe d'identificateurs de ressources est représenté par un objet graphique sur un dispositif d'affichage de l'utilisateur, et ladite signalisation est réalisée au moins par changement d'apparence de cet objet graphique.

L'interface utilisateur

L'utilisateur peut bien sûr directement fournir au système un ensemble d'URI formant une requête de recherche. Néanmoins, d'autres moyens peuvent aussi lui être offerts pour l'assister dans la préparation et la soumission d'une requête de recherche. On va d'abord décrire des moyens qui lui permettent de lancer une requête de recherche à partir d'un seul lien qui se trouve dans une page et qu'il sélectionne très simplement.

Construire une requête de recherche à partir d'un lien dans une page

Pour déclencher l'exécution d'une requête de recherche à partir d'un lien hypertexte se trouvant dans une page, l'utilisateur peut utiliser l'un quelconque des dispositifs parmi les suivants :

- Un objet graphique activable par exemple par clic (e.g. un bouton) est présenté à proximité de certains liens hypertextes (URI) dans une page Web. Son activation déclenche l'envoi d'une requête de recherche contenant l'URI en question.
- Le système est doté d'un moyen apte à basculer la page dans un état où chaque clic sur un lien déclenche l'exécution d'une requête de recherche (contenant ce lien).
- Une touche du clavier, telle que la touche « Ctrl », appuyée alors que l'on clique (par un moyen de pointage) sert à déclencher l'exécution d'une requête de recherche à partir du lien sur lequel curseur du moyen de pointage est positionné.
- Le bouton droit de la souris (ou équivalent) sert à déclencher l'exécution d'une requête de recherche à partir du lien sur lequel le curseur de la souris est positionné.
- Autre dispositif analogue.

Chacun de ces dispositif peut avantageusement permettre d'exécuter ladite requête de recherche en plus de (en parallèle à) l'accès à la page désignée par le lien en question. Le résultat de la requête de recherche sera par exemple affiché dans une deuxième fenêtre (nouvelle instance du navigateur) ou encore dans une sous-fenêtre du navigateur[3].

En supplément du lien sélectionné, d'autres URI peuvent être ajoutés d'office dans la requête de recherche[4]. Ceux-ci peuvent notamment être:

- les liens se trouvant dans la page, dans la région[5] de l'URI sélectionné ;
- les URI précédemment sélectionnés par l'utilisateur pour cette même requête au cours de sa navigation[6] ;
- des liens explicitement prévus et de préférence déterminés par le concepteur de la page pour accompagner l'URI sélectionné ;
- les URI qu'un autre utilisateur (« mentor » ou référent) considère comme étant très pertinents par rapport à l'URI sélectionné, le mentor étant déterminé automatiquement par le système, ou spécifié par l'utilisateur lui-même (choisit dans une liste de « copains » qu'il a au

[3] De manière analogue à la sous-fenêtre existante aujourd'hui pour les liens favoris, cette sous-fenêtre peut être adjacente à la sous-fenêtre principale dans laquelle était affichée la page contenant le lien que l'utilisateur a cliqué et dans laquelle est ensuite affichée la page accédée par le fait de cliquer sur ce lien.

[4] En effet, un des avantages essentiels du système est de pouvoir fonctionner (trouver les ressources d'information pertinentes) même si la requête de recherche est composée d'une pluralité d'URI.

[5] On décrit plus loin un mécanisme pour déterminer la région de pertinence d'un lien dans une page.

[6] Les nouveaux URI trouvés par le système sont alors mis en évidence dans le résultat retourné à l'utilisateur (pour les distinguer des URI qui avaient déjà été retournés dans la même navigation).

préalable mémorisée dans le système), ou encore proposé par le concepteur de la page (l'utilisateur peut aussi choisir dans une liste d' « experts » proposés par le concepteur de la page).

On va maintenant décrire comment l'utilisateur peut préparer une requête composée de plusieurs liens qu'il glane au cours de sa navigation.

Affichage de la requête courante en préparation

Au lieu de déclencher directement une requête de recherche, l'action de l'utilisateur (comme décrit plus haut, par exemple le fait de cliquer sur un lien avec le bouton droit et choisir l'option appropriée) déclenche l'affichage d'une deuxième page dans laquelle :

- en plus du lien que l'utilisateur vient de sélectionner[7], d'autres liens, qu'il a le cas échéant précédemment sélectionnés pour cette même requête, sont présentés ;
    - des cases à cocher peuvent être affichées en association avec chaque lien présenté, de manière à ce que l'utilisateur puisse notamment sélectionner ceux qui vont effectivement former la requête;
- ladite deuxième page est aussi munie d'un moyen d'entrée (tel qu'un bouton) permettant de lancer la requête de recherche.

Ainsi l'utilisateur peut préparer une requête progressivement, en sélectionnant des liens les uns après les autres[8] lors de sa navigation[9] et ensuite envoyer une requête composée de plusieurs URI.

Ladite deuxième page peut en plus contenir des objets graphiques dépliables (comme par exemple des répertoires, casiers, dossiers, ou métaphore analogue) représentant des requêtes en préparation autres que la requête en cours. L'utilisateur peut ainsi choisir la (ou les) requête qui sera enrichie par le nouveau lien qu'il vient de sélectionner.

Suite à la préparation d'une requête à partir d'un URI correspondant à un lien hypertexte dans une page (comme décrit plus haut), les requêtes déjà existantes qui le cas échéant contiennent cet URI lui sont automatiquement présentés en priorité.

Avantageusement, ladite deuxième page peut être composée de deux parties. L'une de ces parties contient les éléments décrits ci-dessus (c'est-à-dire les éléments de la requête en préparation). L'autre partie présente le contenu de la page désignée par le lien sélectionné par l'utilisateur.

Par exemple, si l'utilisateur clique sur un lien alors que la page est à l'état où tous les clics déclenchent l'affichage de la requête courante en préparation (ou avec le bouton droit de la souris, etc), le serveur lui retourne ladite deuxième page qui comprend ainsi :

- dans une partie : les éléments de la requête en préparation
- et dans l'autre partie : le contenu de la page désignée par le lien cliqué.

Ainsi, le fait d'utiliser le système représente un avantage important par rapport à la navigation classique sur le Web : l'utilisateur reçoit non seulement la page désignée par le lien qu'il a cliqué (c'est la navigation classique sur le Web), mais en même temps il bénéficie de la possibilité

---

[7] (ainsi que des liens ajoutés d'office, le cas échéant, comme décrit ci-avant)

[8] (dans une même page ou dans des pages différentes)

[9] (lors d'une même navigation ou de manière plus espacée dans le temps)

d'envoyer une requête (contenant plusieurs URI) pour obtenir encore d'autres ressources pertinentes en relation avec cette page.

D'autres aspects, buts et avantages de la présente invention apparaîtront mieux à la lecture de la description détaillée suivante.

Deux architectures permettant un tel procédé sont présentées aux figures 1a et 1b qui sont explicatives en soi.

En variante, ladite deuxième page est retournée après une exécution rapide (voire restreinte[10]) de la requête de recherche en cours à laquelle le lien cliqué à été ajouté. La deuxième page contient alors directement une partie du résultat[11]. L'utilisateur reçoit alors non seulement la page désignée par le lien qu'il a cliqué, mais en plus il bénéficie directement d'autres ressources pertinentes en relation avec cette page.

Plus avantageusement encore, ladite deuxième page peut être affichée dans une sous-fenêtre[12] adjacente à la sous-fenêtre principale du navigateur. Cette sous-fenêtre adjacente s'ouvre en réponse à l'action de l'utilisateur qui souhaite l'affichage de la requête en préparation (c'est-à-dire ladite deuxième page).[13]

La requête en préparation peut ainsi être affichée en parallèle (de manière asynchrone) à l'affichage de la page désignée par le lien cliqué; cette dernière s'affichant (indépendamment) dans la sous-fenêtre principale.

Le résultat de la requête de recherche peut ensuite être présenté dans la même sous-fenêtre adjacente.

Comme mentionné précédemment, un résultat (partiel) peut éventuellement être retourné après exécution partielle ou restreinte de la requête de recherche en cours, requête à laquelle le lien cliqué à été ajouté. La sous-fenêtre adjacente présente alors directement un résultat rapide de recherche (qui sera éventuellement complété par la suite).

Résultat de l'exécution d'une requête de recherche

Pour chaque requête de recherche, le serveur peut retourner les résultats directement (par exemple en retour de la requête HTTP) ou en différé (par exemple par email).

Le serveur retourne les URI (résultant d'une requête) dans une page présentant la même structure que ladite deuxième page (ou ladite requête en préparation), à savoir :

- des cases à cocher sont associées aux liens de manière à ce que l'utilisateur puisse sélectionner ceux qu'il apprécie et supprimer ceux qu'il n'apprécie pas[14]
    - chaque URI[15] peut ainsi être dans au moins l'un des états suivants[16] : suggéré (état par défaut), accepté ou supprimé (les URI qui sont à l'état supprimé ne sont pas présentés);

[10] Dans le cas d'une requête sur des pages déjà crawlées, le système peut directement retourner les URI (ou pages) pertinents déjà connus et retourner la suite des résultats en différé.

[11] (par exemple sous forme d'une liste d'URI ou un ensemble de vignettes représentant ces pages en miniature)

[12] (analogue à la sous-fenêtre des liens favoris des navigateurs actuels)

[13] Noter que, en parallèle à l'affichage de la requête en préparation, le serveur peut avantageusement déjà commencer à arpenter le Web (crawling en terminologie anglo-saxone) –c'est-à-dire constituer R^-, R^+, R^{+-}, R^+, R^{+-} et R^{+-+} comme décrit par la suite– autour du lien sélectionné.

[14] (c'est-à-dire demander au système de ne plus les suggérer)

- la page est munie d'un moyen d'entrée (tel qu'un bouton) permettant de relancer la requête de recherche.

La page retournée présente également les autres requêtes (du même utilisateur) sous forme d'objets graphiques dépliables, comme déjà décrit. La présentation de celles-ci peut être hiérarchisée selon leur pertinence par rapport au lien cliqué (selon les procédés de calcul de pertinence décrits plus loin).

La page retournée présente des moyens de commande permettant à l'utilisateur de créer de nouvelles requêtes et supprimer des requêtes existantes. Bien entendu, l'utilisateur peut copier-coller des URI à partir de requêtes existantes ou de n'importe quelle autre ressource. Et lorsque le résultat d'une requête est retourné par le serveur, l'utilisateur peut déplacer (ventiler) les URI reçus dans d'autres requêtes. Chaque requête est accessible individuellement au moyen d'un URI qui lui est propre.

Un service de requêtes permanentes (« persistent queries » en terminologie anglo-saxonne) est également mis en œuvre pour rafraîchir (mettre à jour) le contenu des requêtes automatiquement. Les requêtes qui sont rafraîchis changent d'apparence. Les requêtes rafraîchies peuvent également être notifiées par message électronique. Le système se prête ainsi bien à une activité de veille sur le Web (surveillance de l'apparition de nouvelles ressources d'information sur le Web).

Le procédé de base

Selon un premier aspect de l'invention, on prévoit un procédé de calcul de score de pertinence d'un URI par rapport à un ensemble d'URI donné en entrée (pour constituer ladite requête). Ce procédé est basé sur une analyse des liens hypertextes et ne nécessite pas d'analyser le contenu des pages pointées, ni le texte autour des liens hypertextes. Optionnellement, divers procédés de catégorisation permettent notamment de garder en mémoire certains résultats de calcul et de ne pas les effectuer à nouveau.

L'idée essentielle du calcul du score de pertinence (d'une page P_2 par rapport à une page donnée P_1) est la suivante[17] :

Soit p_1 la probabilité[18] qu'un auteur aléatoire (de page Web) mette dans une page un lien sur P_1.

Soit p_2 la probabilité qu'un auteur aléatoire mette dans une page un lien sur P_2.

Soit p_{12} la probabilité qu'un auteur aléatoire, qui a mis un lien sur P_1 dans une page, mette un lien sur P_2 dans la même page (ou vice-versa). Si P_2 est pertinente par rapport P_1, alors p_{12} est supérieur au produit $p_1.p_2$ (en effet, le fait de trouver P_1 intéressant augmente les chances de trouver P_2 intéressant, et vice-versa). On pourrait ainsi considérer que le score de pertinence est simplement le rapport $p_{12}/p_1.p_2$ Cependant avec cette approche on ne peut pas traiter le cas de plus de deux pages, du moins pas sans perte d'information. Alors on procède plutôt comme suit :

---

[15] Optionnellement, la présentation du résultat d'une requête de recherche inclut le contenu des pages (pointées par les URI résultants) par exemple sous forme miniaturisée (vignettes).

[16] Accessoirement, une possibilité de copie (« gel ») de page (en local ou dans un espace personnel sur un serveur) peut aussi être offert à l'utilisateur. Chaque lien peut alors être dans un des états suivants : suggéré, accepté, supprimé ou gelé.

[17] Ci-après, nous allons considérer que P_1 et P_2, (ou P_i, P_j etc) sont des pages Web, bien que les procédés décrits soient bien plus généraux, comme on l'a déjà mentionné brièvement.

[18] La probabilité d'être intéressé par une (ou certaines) page(s) est approchée en comptant le nombre de pages qui ont un lien sur elle(s) et en divisant ce nombre par une estimation du nombre de pages qui auraient pu en avoir.

La pertinence d'une page **par rapport à un ensemble de pages** peut être définie par la « quantité de raisons communes » d'être intéressé par toutes ces pages.

Des calculs algébriques permettent d'obtenir des équations donnant la quantité de raisons communes entre plusieurs pages. Cette quantité (ou proximité, ou encore homogénéité) est notée x, avec en indice les pages dont il est question ; la probabilité d'être lié à une certaine page P_i est notée p_i ; la probabilité d'être lié à *au moins* une page parmi P_i, P_j, …, P_n est notée $p_{ij...n}$. :

$$\overline{x_{ij}} = \frac{\overline{p_i} \cdot \overline{p_j}}{p_\emptyset \cdot p_{ij}},\ \overline{x_{ijk}} = \frac{\overline{p_i} \cdot \overline{p_j} \cdot \overline{p_k} \cdot \overline{p_{ijk}}}{p_\emptyset \cdot p_{ij} \cdot p_{ik} \cdot p_{jk}},$$

et ainsi de suite (tous les sous-ensembles impairs au numérateur, et les autres au dénominateur)[19,20].

Ainsi le procédé est le suivant :

Soit R l'ensemble de pages de la requête, et R_X (avec $R_X \cap R = \emptyset$) un ensemble de pages que l'utilisateur refuse explicitement.

Cadre général pour le traitement par l'amont

Nous supposons ici que R est « homogène par l'amont » – nous décrirons ce terme plus loin (voir la section « Décomposer une requête en sous-requêtes homogènes »).

Procédure :

1. Pour chaque page P_i de R, trouver $B(P_i)$, l'ensemble des pages pointant sur elle, et trouver R^-, l'ensemble des pages pointant sur au moins une des pages de R :
   $R^- = \bigcup_{P_i \in R} B(P_i)$.
2. Pour chaque page P_i de R^-, trouver $F(P_i)$, l'ensemble des pages pointées par P_i, et trouver R^{-+}, l'ensemble des pages pointées par au moins une page pointant sur une page de R (noter que $R \subset R^{-+}$) :
   $R^{-+} = \bigcup_{P_i \in R^-} F(P_i)$
3. Pour chaque page P_i de R^{-+}, trouver $B(P_i)$, l'ensemble des pages pointant sur elle, et trouver R^{-+-}, l'ensemble des pages pointant sur au moins une des pages de R^{-+} (noter que $R^- \subset R^{-+-}$) :
   $R^{-+-} = \bigcup_{P_i \in R^{-+}} B(P_i)$
4. Exécuter une procédure de calcul de pertinence par l'amont avec :
   - H = un sous-ensemble de R^{-+} contenant au moins R^-
   - A=R^{-+}
   - K=R
   - T=R_X.

[19] Les barres supérieures indiquent des compléments, et $p_\emptyset$, la probabilité d'aimer au moins une page d'un ensemble vide, est une constante égale à zéro ; elle est présente dans l'équation pour des raisons de cohérence.

[20] Notons que les deux méthodes sont équivalentes pour le cas de deux pages (ci-dessous nous notons u_{12} la probabilité d'être lié à P_1 ou P_2, et p_{12} la probabilité d'être lié à P_1 et P_2) :

$$\left(\frac{\overline{p_1} \cdot \overline{p_2}}{u_{12}}\right) > 0 \quad \Leftrightarrow \quad \frac{p_{12}}{p_1 \cdot p_2} > 1$$

Cadre général d'un procédé de calcul de pertinence par l'amont

Entrée:

- Un ensemble H (comme "Hub" qui signifie "pivot") de pages hub (« pages citantes »),
- Un ensemble A (comme "Authority") de pages autorités candidates (« pages citées »),
- Un ensemble K (comme "Kernel") de pages de référence (auxquelles les pages de A sont à comparer), $K \subset A$,
- Et optionnellement un ensemble T (comme "Trash") de pages que l'utilisateur déclare explicitement ne pas vouloir, $T \cap K = \varnothing$.

Sortie: La procédure va retourner un score de pertinence r (comme "relevancy") pour chaque page de A.

Procédé de base

Pour chaque page P_i de A calculer son score de pertinence r_i, qui est égal à $x_{i \cup K}$ défini comme suit.

L'« homogénéité » x_S d'un ensemble S (ou la « proximité » de ses éléments) est défini comme suit (c'est l'équation de quantité de raison introduite précédemment):

$\overline{x_S} = \prod_{P \subset S} \overline{b_P}^{\sigma_P}$, où

- $\sigma_P = \begin{cases} -1 \text{ si P contient un nombre pair de pages} \\ +1 \text{ sinon} \end{cases}$
- $b_P = \dfrac{\left| \bigcup_{P_i \in P} B(P_i) \right|}{|H| + \hbar}$

  où $\hbar$ est une valeur[21] égale par exemple à une constante (dans l'exemple présenté ci-après, nous prenons $\hbar = 1$), à $|H|$, à $|A|$, ou encore à une combinaison de ces derniers ; $\hbar$ peut avantageusement être une fonction de la popularité des pages $P_i \in P$ et/ou leur date d'apparition sur le Web. Noter que pour éviter le cas pathologique de division par zéro[22] il faut que $\hbar \neq 0$.

- autrement dit[23] : $b_P = \dfrac{\sum_j l_{jP}}{|H| + \hbar}$ où $\sum_j l_{jP}$ est le nombre de pages (de H) qui pointent sur au moins une page de P (P étant le sous-ensemble courant de S qui est considéré)

  (i.e. $l_{jP} = \begin{cases} +1 \text{ si } \exists P_i \in P \mid l_{ji} = +1 \\ 0 \text{ sinon} \end{cases}$ et[24] $l_{ji} = \begin{cases} 0 \text{ s'il n'y a pas de lien entre } P_j \text{ et } P_i \\ 1 \text{ s'il y a un lien entre } P_j \text{ et } P_i \end{cases}$;

---

[21] En principe, $|H| + \hbar$ est censée être une approximation du nombre de pages dont les auteurs connaissent l'existence de la page (ou de l'ensemble des pages) en question ; ainsi si dans ces pages un lien n'a pas été mis ce n'est pas par méconnaissance de la (ou des) page(s) en question mais volontairement.

[22] Si $\hbar = 0$, dans le cas où au dénominateur la (ou l'une des) union(s) $\bigcup_{P_i \in P} B(P_i)$ couvre toutes les pages de H, le rapport entre la cardinalité de cette union et $|H| + \hbar$ donnerait 1 et son complément donnerait 0 ; on aurait alors une division par zéro.

[23] Cette autre définition est donnée pour conformité avec la suite.

en d'autres termes, l_{ji} est égal à 1 s'il y a un lien

- d'une page P_j (de H)
- à au moins une page P_i de P

et zéro sinon).

Exemple

On va maintenant illustrer ce procédé par l'exemple présenté à la figure 2, dans lequel :

- la requête K (c'est-à-dire R) est composée d'une seule page, la page 0 ;
- R est composée des pages 9, 10, 11, 12 et 13 ;
- A (c'est-à-dire R^*) est l'ensemble des pages 0, 1, 2, 3 et 4 ;
- $H=R^+$ est l'ensemble des pages 5, 6, 7, 8, 9, 10, 11, 12 et 13 ;
- T est vide.

Calcul du score de pertinence de chaque page de A

Prenons $\hbar = 1$, comme $|H| = 9$ on a $|H| + \hbar = 10$.

Pour la page 1, le score de pertinence est :

$r_1^+ = 1-(1 - x)*(1 - y) / (1.0 - z) = 0.2$

puisque

x=0.5 est le nombre sur $|H| + \hbar$ des pages citant la page 0, qui sont: 9,10,11,12,13

1, qui sont: 10,11

y=0.2

z=0.5

0 ou 1, qui sont: 9,10,11,12,13

Pour la page 2

$r_2^+ = 1-(1 - x)*(1 - y) / (1 - z) = -1$

puisque

x=0.5 est le nombre sur $|H| + \hbar$ des pages citant la page 0, qui sont: 9,10,11,12,13

2, 5,6,7,8,12,13

y=0.6

z=0.9

0 ou 2, 5,6,7,8,9,10,11,12,13

La page 2 est citée par beaucoup de pages qui ne citent pas la page 0. C'est la raison pour laquelle son score est négatif.

Pour la page 3

$r_3^+ = 1-(1 - x)*(1 - y) / (1 - z) = 0.1$

puisque

x=0.5 est le nombre sur $|H| + \hbar$ des pages citant la page 0, qui sont: 9,10,11,12,13

3, 10

y=0.1

z=0.5

0 ou 3 9,10,11,12,13

Pour la page 4

$r_4^+ = 1-(1 - x)*(1 - y) / (1 - z) = 0.1$

[24] Noter que le cas $l_{ji} = 0$ n'est là que pour la cohérence dans la formalisation mathématique. En pratique, on ne considère que les liens existants, le cas $l_{ij}=0$ n'apparaît donc pas.

. puisque

x=0.5	est le nombre sur $\|H\|+\hbar$ des pages citant la page	0, qui sont:	9,10,11,12,13
y=0.1		4,	9
z=0.5		0 ou 4,	9,10,11,12,13

Enfin, pour la page 0

$r_4^+ = 1-(1-x)*(1-x)/(1-x) = 0.5$

puisque

x=0.5	est le nombre sur $\|H\|+\hbar$ des pages citant la page	0, qui sont:	9,10,11,12,13

Filtrer le « Népotisme »

Pour chaque lien l_{ji} existant on peut lui associer un poids en fonction de la proximité des pages P_i et P_j et améliorer ainsi le résultat. On appelle cela « filtrer le Népotisme » car cela permet de diminuer l'importance des liens entre pages qui se promeuvent mutuellement. Typiquement on arrive ainsi à filtrer par exemple les liens des « sommaires » et autres « menus » qui, de manière répétitive, se trouvent dans toutes les pages d'un site.

L'idée de base consiste à affaiblir les liens reliant deux pages que nous savons proches, en affectant un poids à chaque lien, poids qui sera égal au complément de la proximité des deux pages reliées (plus la proximité est grande, plus le lien doit être affaibli). Une fois que les liens sont ainsi pondérés, il est possible de calculer l'homogénéité d'un ensemble de pages en utilisant non plus le nombre de pages citantes, mais la somme de leurs poids.

Dans le Procédé de base, on remplace ainsi la définition de $\sum_j l_{jP}$ par $\sum_j \ell_{jP}$

où $\ell_{jP} = \begin{cases} min\left[1; \underset{i \in P}{\text{Max}}\ \overline{x_{ji}}\right] \text{ si } \exists P_i \in P \mid l_{ji} = +1 \\ 0 \text{ sinon} \end{cases}$

Explications :

- $\underset{i \in P}{\text{Max}}\ \overline{x_{ji}}$ est le complément de la proximité entre la page $P_j \in H$ en question et la page $P_i \in P$ pour laquelle le lien entre P_j et P_i présente la proximité minimum
- $min\left[1; \underset{i \in P}{\text{Max}}\ \overline{x_{ji}}\right]$ signifie que cette valeur est tronquée supérieurement à 1
- et toujours $l_{ji} = \begin{cases} 0 \text{ s'il n'y a pas de lien entre } P_j \text{ et } P_i \\ 1 \text{ s'il y a un lien entre } P_j \text{ et } P_i \end{cases}$

En d'autres termes, s'il y a au moins un lien entre

- la page P_j (de H) en question
- et une page P_i de P,

l_{jP} est égal au complément de la proximité entre la page P_j et la page P_i qui lui est la moins proche et vers laquelle elle possède un lien. $\sum_j \ell_{jP}$ est la somme des poids ainsi associés aux pages de H qui pointent sur au moins une des pages du sous-ensemble P considéré.

Pour déterminer la proximité x_{AB}, on peut prendre l'équation de quantité de raisons communes (déjà décrite) : $\overline{x_{AB}} = \frac{\overline{p_A} \cdot \overline{p_B}}{p_\phi \cdot p_{AB}}$

Exemple numérique:
La figure 3 présente un exemple où le nombre de pages pointant sur la page A est égal à 0.9+0.2+0.4+0.8=2.3
Le nombre de pages pointant sur la page B est égal à 0.9+0.1+0.3+0.5=1.8
Le nombre de pages pointant sur A ou B (N p_{AB}) est égal à 0.9+0.2+0.9+0.8+0.3+0.5=3.6
Ainsi, si on considère que $|H| + \hbar = 100$, le calcul de la proximité de A et B donne :

$\overline{x_{AB}} = \frac{\overline{p_A} \cdot \overline{p_B}}{p_\phi \cdot p_{AB}} = \frac{0.977 \cdot 0.982}{1 \cdot 0.964}$, ce qui donne $\tilde{x}_{AB} = \frac{x_{AB}}{p_B} \approx 0.264 = 26.4\%$.

Alternativement (puisqu'on doit calculer la proximité entre deux pages seulement), on peut aussi prendre n'importe quelle autre équation donnant la proximité entre deux pages, telle que par exemple l'équation préférée suivante, que nous adoptons dans la suite pour calculer le facteur de népotisme:

$$x_{ij} = \frac{|B(P_i) \cap B(P_j)|}{|B(P_i) \cup B(P_j)|} \cdot \frac{\min[|B(P_i)|;|B(P_j)|]}{\max[|B(P_i)|;|B(P_j)|]}$$

On peut compléter le procédé de base en lui ajoutant un procédé itératif, dit de « distillation », permettant d'associer aux pages des scores de qualité et popularité et de trouver ainsi des URI pertinents supplémentaires en réponse à la requête.

Ce procédé de distillation est dérivé de l'idée originale de Jon Kleinberg [(« Method and system for identifying authoritative information resources in an environment with content-based links between information resources », brevet US-A-6 112 202, reprise par d'autres chercheurs) selon laquelle chaque page peut être caractérisée par un score « pivot » (score « hub » en terminologie anglosaxonne) et un score « autorité ». Respectivement, ces scores quantifient la *qualité de citation* et la *popularité* d'une page. Le score autorité d'une page est obtenu à partir de la somme des scores hub des pages ayant un lien sur elle (au début on leur donne une valeur unitaire). Symétriquement, le score hub d'une page est obtenu à partir du total des scores autorités des pages qu'elle pointe. Pour calculer ces scores, on procède itérativement en normalisant (ou en mettant à l'échelle) les scores à chaque itération. Le procédé converge après un petit nombre d'itérations.

La solution proposée

Les approches de distillation[25] permettent de trier ou sélectionner des pages selon un critère de qualité « dans l'absolu »[26]. La présente invention permet, au contraire, de trier ou sélectionner des pages selon un critère de pertinence par rapport à une requête donnée.

Pour faciliter la lecture, on va considérer les 7 ensembles (voir la figure 4) suivants:

- R est constitué par les pages de la requête (ou plus exactement leurs URI) et on suppose ici que ces dernières sont homogènes au sens du procédé de base[27].
- R^- est l'ensemble des pages qui contiennent un lien vers[28] au moins une des pages de la requête.
- R^{-+} est l'ensemble des pages pointées (citées) par les pages R^-.
- R^{-+-} est l'ensemble des pages qui citent les pages R^{-+} ($R^- \subset R^{-+-}$).
- R^+ est l'ensemble des pages citées par au moins une des pages de la requête (R).
- R^{+-} est l'ensemble des pages qui citent les pages R^+.
- R^{+-+} est l'ensemble des pages citées par les pages de R^{+-} ($R^+ \subset R^{+-+}$).

Procédé par l'amont

L'idée de base

Les probabilités dont il est question dans le procédé de base font intervenir le nombre (le comptage) des pages de R^- qui contiennent un lien donné ou un lien parmi un ensemble de liens donnés (vers des pages de R^{-+}). On gagnerait à pondérer ce nombre par la *qualité de citation* (score hub) de chaque page qui contient un tel lien.

On voudrait ainsi qu'une page de R^- citant plus de meilleures pages (de R^{-+}) soit considérée comme étant de meilleure qualité de citation, et qu'en retour un poids plus fort lui soit donné dans le cadre du calcul des scores[29] des pages qu'elle cite (R^{-+}), les scores des pages de R^- et ceux des pages de R^{-+} s'influençant mutuellement dans une approche itérative (de renforcement bipartite) qui converge (pour les mêmes raisons que pour la distillation[30]).

Extension R^{-+-}

Le nombre de pages de R^{-+-} citant chaque page candidate (de R^{-+}) intervient aussi dans les calculs. Or leur prise en compte coûte cher. On va alors approximer les résultats en ne considérant que

---

[25] (ou les approches analogues telles que PageRank [Lawrence Page, « method for node ranking in a linked database », US Pat. 6,285,999])

[26] Ainsi, selon ces approches, une meilleure page est soit une page plus populaire soit une page qui cite plus de pages plus populaires.

[27] On verra plus loin comment décomposer une requête en sous-requêtes homogènes.

[28] (autrement dit « qui citent », ou encore « qui pointent »)

[29] Rappelons qu'il s'agit ici de scores de pertinence par rapport à la requête, contrairement à l'approche de distillation qui permettait de déterminer un score de qualité « dans l'absolu ».

[30] Notons seulement que, contrairement aux scores autorité de l'approche originale de distillation, le calcul du score de pertinence d'une page de R-+ peut résulter en une valeur négative (que l'on va alors neutraliser ; ceci est décrit plus loin). En effet, certaines pages peuvent être, non seulement pas proches de la requête, mais même antagonistes par rapport à elle (le fait d'y être intéressé diminue les chances d'aimer les pages de la requête et inversément).

celles qui citent les pages candidates ayant un bon score, ce score étant calculé d'abord en ne considérant que R et ensuite en étendant cet ensemble vers R^+ progressivement.

Trouver les pages récentes

Pour calculer le score de pertinence d'une page candidate, au lieu de prendre le résultat de l'équation de quantité de raisons (voir, dans la description du procédé de base, l'équation $\overline{x_S} = \prod_{P \subset S} \overline{b_P}^{\sigma_P}$) directement, il est préférable de multiplier ce résultat x_S par le score autorité de la page candidate en question (simplement calculé à partir du total des scores hub des pages citantes), afin d'affaiblir ainsi les pages qui sont relativement moins fiables (car moins populaires). Les scores hub des pages citantes (R^+) ainsi trouvés vont ensuite servir à déterminer les pages qui sont pertinentes sans être populaires, au moyen d'une équation de proximité, par exemple $x_{ij} = \frac{|B(P_i) \cap B(P_j)|}{|B(P_i) \cup B(P_j)|} \cdot \frac{\min[|B(P_i)|;|B(P_j)|]}{\max[|B(P_i)|;|B(P_j)|]}$[31], dans laquelle les cardinalités d'ensemble (représentées entre barres verticales) vont être remplacées par le total des scores hub des pages en question[32].

En résumé : d'abord on détermine les pages pertinentes (et populaires à la fois) par la première méthode qui en plus donne les scores hub des pages citantes (et de manière fiable). Ces scores permettent ensuite de trouver les pages pertinentes mais pas (encore) populaires.

Népotisme avec scores hub

Le filtrage du népotisme dans le procédé de base utilise un facteur de népotisme $\overline{x_{ji}}$. Puisque nous avons maintenant les scores[33] des pages citantes, nous pouvons améliorer le procédé en prenant $\overline{x_{ji} \cdot h_j}$ comme facteur de népotisme (au lieu de $\overline{x_{ji}}$), où h_j est le score de la page citante. En effet, il faut d'autant plus affaiblir un lien népotiste (d'une page citante P_j à une page citée P_i) que le score de la page citante P_j est faible.

Régions de pertinence dans les pages

Après une première itération, dans les pages citantes le système peut

- repérer les régions contenant des liens dirigés sur des pages de R^+ ayant un bon score
- et commencer déjà à élaguer les liens qui ne sont pas situés dans ces régions.

Comme les liens en question se trouvent placés sous des nœuds d'une structure typiquement arborescente de document (tel qu'en HTML notamment), pour déterminer une région de pertinence il suffit de prendre les nœuds (minimaux) qui englobent tous les bons liens et de leur retrancher les sous-nœuds (maximaux) qui contiennent un mauvais lien et qui ne contiennent pas de bon lien. Par « bon » lien on entend : un lien dirigé sur une page ayant un bon score. Par « mauvais » lien, on entend : un lien qui a été explicitement refusé par l'utilisateur.

---

[31] Cette équation donne la proximité entre une page candidate et une page de la requête. Pour trouver la proximité d'une page candidate avec l'ensemble des pages de la requête, on peut prendre la moyenne arithmétique des proximités de la page candidate avec chaque page de la requête.

[32] On peut dire que l'on remplace les cardinalités par des « cardinalités pondérés », les poids étant les scores hub.

[33] (que ce soit de manière absolue ou par rapport à la requête)

Procédé par l'aval

Toutefois, dans le cas de pages de R qui sont fortement « hub » (qui contiennent un nombre significatif de liens sur des pages de R^+) mais trop faiblement « autorité » (il y a trop peu de pages de R^- qui les citent), le procédé de calcul de pertinence par l'amont ne peut donner de résultat satisfaisant. En effet, il n'y a pas un nombre significatif de liens entrants sur lesquels se baser.

On utilise alors le même procédé exactement mais de manière symétrique (voir la figure 4): les itérations se font entre R^{+-} et R^+ (ou R^{+-+}) au lieu de entre R^{-+} et R^- (ou R^{-+-}).

Ainsi, on forme l'ensemble R^+ et on considère comme candidates les autres pages qui les citent (R^{+-}). Les comptages sont pondérés : les pages de R^+ (ou R^{+-+}) citées par plus de meilleures pages (de R^{+-}) ont un poids plus fort dans le calcul des scores des pages candidates (de R^{+-}), c'est le renforcement bipartite déjà décrit plus haut.

Attribution de pages « artificielles »

Les procédés par l'amont et par l'aval peuvent être avantageusement intégrés de la manière suivante :

Après le traitement par l'amont (éventuellement même après chaque itération amont), aux pages candidates (R^{-+}) ayant obtenu un score de pertinence suffisant, on associe à l'aval un ensemble de pages supplémentaires (« pages artificielles ») dont la cardinalité est fonction dudit score de pertinence. Chaque page artificielle est aussi citée par (au moins) une page de la requête. On donne ainsi à l'aval un « avantage » aux scores de ces bonnes pages (de R^{-+}) trouvées par l'amont[34], et par conséquent on donne aussi indirectement un avantage aux scores des pages (de R^{+-+}) citées le cas échéant par ces bonnes pages.

Et réciproquement, après le traitement par l'aval (éventuellement même après chaque itération aval), on applique à l'amont le même procédé de manière symétrique. On favorise ainsi les bonnes pages de R^{+-} et indirectement les pages (de R^{-+-}) qui les citent le cas échéant.

Le fait de ne pas amalgamer les scores par l'amont (des pages R^{-+}) avec les scores par l'aval (pages R^{+-}) permet de les dissocier dans les calculs. Notamment, on peut diminuer l'influence des scores obtenus par l'aval[35] dans les traitements par l'amont ou vice-versa.

Par ailleurs, grâce à cette idée de « pages artificielles », la présente invention peut être appliquée en complément à l'approche plus classique de distillation (ou procédé analogue tel que « PageRank » par Lawrence Page, « Method for node ranking in a linked database », voir US-A-6 285 999). En effet, une fois les scores de pertinence obtenus pour chaque page, on peut

---

[34] Noter que, avantageusement, ceci se fait sans amalgamer les scores de pertinence par l'amont et par l'aval.

[35] Dans le cas où les scores par l'aval sont jugés moins fiables que ceux par l'amont, on pourra multiplier chaque score obtenu par l'aval par une constante ou une fonction de fiabilité ou crédibilité,

- avant de le prendre en compte à l'amont
- avant de calculer éventuellement un score global (amont-aval) à retourner à l'utilisateur.

Cette fonction de fiabilité ou crédibilité peut être basée sur le procédé décrit dans le dépôt de brevet PCT/FR00/03157.

Noter aussi que les pages citantes (R^-) ayant de très bons scores en résultat du traitement par l'amont, et les pages citées (R^+) ayant de très bons scores en résultat du traitement par l'aval, pourront aussi être insérés dans la réponse fournie à l'utilisateur.

modifier artificiellement les nombres respectifs des pages citantes et citées avant de lancer la distillation (ou procédé analogue).

On comprend ainsi qu'il est très avantageux d'appliquer en premier le nouveau procédé de calcul de pertinence introduit ici, et ajouter des pages artificielles comme décrit ci-dessus, avant d'appliquer l'approche originale de distillation (ou analogue) sur l'ensemble qui en résulte. En effet, on peut alors arpenter (« crawling » en terminologie anglo-saxonne) le Web en suivant les liens (en amont et en aval) autour des pages des 7 ensembles précédemment citées, en tirant parti de l'ajout des pages artificielles pour avantager les pages Web liées aux pages qui sont plus pertinentes par rapport à la requête.

Arpenter le Web récursivement

Dans la mesure où les pages ayant les meilleurs scores sont censées être très pertinentes pour l'utilisateur (et dans la mesure où la pertinence est transitive), les procédés décrits ici pourront être récursivement appliqués sur ces dernières pour découvrir encore d'autres pages pertinentes. On peut ainsi arpenter le Web à partir de la requête de l'utilisateur.

La figure 5 présente de manière schématique un tel procédé : la recherche de pages pertinentes peut être appliquée récursivement sur les ensembles « Bonnes pages trouvées par l'amont », « Bonnes pages trouvées par l'aval », « Bonnes pages hub » et « Bonnes pages autorités » qui dans la figure 5 sont encadrés par des rectangles. A chaque récursion, les scores des meilleures pages trouvées deviennent un peu plus faibles et le procédé s'arrête quand les scores cessent d'être suffisants.

Prise en compte des requêtes

Les requêtes des utilisateurs (qui sont des ensemble d'URI) peuvent être exploitées au même titre que les pages Web.

Description plus détaillée

On calcule le score de pertinence par l'amont comme suit:

Nouveau procédé de calcul de score de pertinence par l'amont

1. Associer à chaque page P_i de H, un nombre h_i, mis initialement à $\frac{1}{|H|+\hbar}$, son score hub[36].

2. (Re-)calculer les scores de pertinence :
   a. Pour chaque page P_i de A calculer r_i^+, égal à $w_{i\cup K}$

      $r_i^+ = w_{i\cup K}$

[36] Ainsi, avantageusement, la sommes des scores h_i plus la somme des scores des $\hbar$ « pages virtuelles », chaque page virtuelle ayant un score de $\frac{\hbar}{|H|+\hbar}$, est égale à 1.

et dans le cas où le résultat est négatif (cas d'une page antagoniste à *R*) neutraliser les liens entrants de manière à avoir $r_i^+ = 0$.

L'homogénéité par l'amont w_S d'un ensemble *S* est défini comme suit:

$\overline{w_S} = \prod_{P \subset S} \overline{a_P}^{\sigma_P}$, où

- $\sigma_P = \begin{cases} -1 \text{ si P contient un nombre pair de pages} \\ +1 \text{ sinon} \end{cases}$
- $a_P = \sum_j h_j l_{jP}$ où

$$l_{jP} = \begin{cases} +1 \text{ si } \exists P_i \in P \mid l_{ji} = +1 \\ 0 \text{ sinon} \end{cases},$$

avec $l_{ji} = \begin{cases} 0 \text{ s'il n'y a pas de lien entre } P_j \text{ et } P_i \\ 1 \text{ s'il y a un lien entre } P_j \text{ et } P_i \end{cases}$

En d'autres termes, l_{jP} est égal à 1 s'il y a un lien

- d'une page P_j (de *H*)
- à au moins une page P_i de *P*

et zéro sinon.

Ceci signifie tout simplement que a_P est le total des scores hub des pages (de *H*) qui pointent sur au moins une page de *P* (*P* étant le sous-ensemble courant de *S* qui est considéré).

Pour chaque lien l_{ji} existant on peut lui associer un poids en fonction de la proximité des pages P_i et P_j et améliorer ainsi le résultat - voir plus loin « Filtrer le Népotisme ».

Ici, puisque $\forall P_i \in K$ on a $r_i^+ = w_K$ (la pertinence est la même pour toutes les pages P_i de *K*), le score de pertinence r_i^+ ne doit être calculée qu'une seule fois pour les pages de *K* (elle sera d'ailleurs déjà calculée lors de la procédure de découpage de la requête *R* en sous-requêtes *K*, et sera donc déjà connue à l'entrée de la procédure).

b. (Ce point sera sauté la première fois.) Pour avoir leur somme égal à 1, on doit diviser chaque r_i^+ par la somme $\sum |r_i^+|$ de toutes les valeurs absolues des $\mathbf{r_i}^+$.

Soit $\delta = \sum_i \left| r_i - \frac{r_i^+}{\sum |r_i^+|} \right|$, la variation globale du score de pertinence.

Si $\delta < \varepsilon$ ($\varepsilon > 0$ étant une marge d'erreur), on considère avoir convergé et le procédé s'arrête. Sinon, le procédé continue.

c. On remplace $\mathbf{r_i}$ par $\frac{r_i^+}{\sum_i |r_i^+|}$

$$r_i \mapsto \frac{r_i^+}{\sum_i |r_i^+|}$$

on peut aussi utiliser un facteur de frottement τ :

$$r_i \mapsto \tau r_i + \bar{\tau}\frac{r_i^+}{\sum_i |r_i^+|} . (\tau \in]0;1[)$$

3. [37]Pour chaque page P_i de H :
   a. Trouver tous les liens qui pointent sur une page ayant un score de pertinence plus grand qu'un seuil epsilon à choisir ($\varepsilon>0$).
   b. Trouver I_i, le plus petit élément HTML [38] contenant la totalité des liens trouvés au point a ci-dessus.
   c. Pour chaque lien pointant sur une page de T (si T n'est pas vide), trouver le plus grand élément HTML le contenant (s'il y en a) et ne contenant pas de lien trouvé au point a. ci-dessus, et l'enlever de I_i.
   d. On garde tous les liens restant dans I_i et on supprime les autres (ou bien on les neutralise en mettant leur l_{ij} à zéro)

4. Recalculer les scores hub:
   a. Pour chaque page P_i de H, calculer $h_i^+ = \sum_j l_{ij} r_j$, la somme des scores de pertinence des pages pointées.
   b. $h_i \mapsto \frac{h_i^+}{\sum |h_i^+|} \cdot \frac{|H|}{|H|+\hbar}$

      Explications :
      - La division par $\sum |h_i^+|$ est, comme pour le score de pertinence, pour garder leur somme égale à 1.
      - Comme on a toujours au moins une page virtuelle en plus des pages de l'ensemble H, on remet à l'échelle en multipliant par $\frac{|H|}{|H|+\hbar}$.

      Ou avec un facteur de frottement τ :

$$h_i \mapsto \tau h_i + \bar{\tau}\left(\frac{h_i^+}{\sum |h_i^+|} \cdot \frac{|H|}{|H|+\hbar}\right) \text{ avec } \tau \in]0;1[$$

Ensuite retourner au point 2.

[37] Ce point peut éventuellement être ignoré après la première fois.

[38] (ou autre représentation analogue…)

Exemple

On va maintenant illustrer le procédé par le même exemple déjà présenté et illustré à la figure 2.

Prenons $\hbar = 1$.

1. *Initialisation des scores hub*

Les scores hub de chaque page de H sont initialisés à $\frac{1}{|H|+1} = 0.1$

2. *Calcul des scores de pertinence*

Pour la page 1, le score de pertinence non encore mis à l'échelle est :
$r_1^+ = 1-(1 - x)*(1 - y) / (1.0 - z) = 0.2$
puisque
x=0.5 --- somme des scores hub des pages citant la page 0, qui sont: 9,10,11,12,13
y=0.2 --- somme des scores hub des pages citant la page 1, qui sont: 10,11
z=0.5 --- somme des scores hub des pages citant 0 ou 1, qui sont: 9,10,11,12,13

Pour la page 2
$r_2^+ = 1-(1 - x)*(1 - y) / (1 - z) = -1$
puisque
x=0.5 --- somme des scores hub des pages citant 0, qui sont : 9,10,11,12,13
y=0.6 --- somme des scores hub des pages citant 2, qui sont : 5,6,7,8,12,13
z=0.9 --- somme des scores hub des pages citant 0 ou 2, qui sont : 5,6,7,8,9,10,11,12,13
La page 2 est citée par beaucoup de pages qui ne citent pas la page 0. C'est la raison pour laquelle son score est négatif. On va donc le neutraliser : on supprime les liens dirigés vers 2.
La même équation donne alors $r_2^+ = 0$

Pour la page 3
$r_3^+ = 1-(1 - x)*(1 - y) / (1 - z) = 0.1$
puisque
x=0.5 --- somme des scores hub des pages citant 0, qui sont : 9,10,11,12,13
y=0.1 --- somme des scores hub des pages citant 3, qui sont : 10
z=0.5 --- somme des scores hub des pages citant 0 ou 3, qui sont : 9,10,11,12,13

Pour la page 4
$r_4^+ = 1-(1 - x)*(1 - y) / (1 - z) = 0.1$
puisque
x=0.5 --- somme des scores hub des pages citant 0, qui sont : 9,10,11,12,13
y=0.1 --- somme des scores hub des pages citant 4, qui sont : 9
z=0.5 --- somme des scores hub des pages citant 0 ou 4, qui sont : 9,10,11,12,13

Enfin, pour la page 0
$r_4^+ = 1-(1 - x)*(1 - x) / (1 - x) = 0.5$
puisque
x=0.5 --- somme des scores hub des pages citant 0, qui sont : 9,10,11,12,13

La somme des r_i^+ est 0,9. Ainsi,

r_0=0.5555
r_1 = 0.2222
r_3= 0.1111
r_4=0.1111
r_2=0

3. *Filtrer dans les pages les régions non pertinentes*

Dans cet exemple T est vide et les pages ne contiennent aucun autre lien que ceux indiqués.

4. *Re-calcul des scores hub*

$h_9^+ = 0.5555 + 0.1111 = 0.6666$
$h_{10}^+ = 0.5555 + 0.2222 + 0.1111 = 0.8888$
$h_{11}^+ = 0.5555 + 0.2222 = 0.7777$
$h_{12}^+ = 0.5555 + 0 = 0.5555$
$h_{13}^+ = 0.5555 + 0 = 0.5555$

Les scores hub mis à l'échelle ($h_i \mapsto \frac{h_i^+}{\sum |h_i^+|} \cdot \frac{9}{10}$) sont ainsi les suivants :

h9 = 0,1742
h10 = 0,2323
h11 = 0,2032
h12 = 0,1452
h13 = 0,1452

et comme le score r de la page 2 est nul, les autres pages de R^+ ont aussi un score nul :

h5 = 0
h6 = 0
h7 = 0
h8 = 0

2. *Calcul des scores de pertinence* (2e itération)

Pour la page 1, le score de pertinence non encore mis à l'échelle est maintenant:
r_1^+=1-(1 - x)*(1 - y) / (1.0 - z) = 0.4355
puisque
x=0.9 --- somme des scores hub des pages citant la page 0, qui sont: 9,10,11,12,13
y=0.4355 --- somme des scores hub des pages citant la page 1, qui sont: 10,11
z=0.9 --- somme des scores hub des pages citant 0 ou 1, qui sont: 9,10,11,12,13

Pour la page 2
r_2^+=1-(1 - x)*(1 - y) / (1 - z) = 0
puisque
x=0.9 --- somme des scores hub des pages citant 0, qui sont : 9,10,11,12,13
y=0 --- somme des scores hub des pages citant 2, qui sont :
z=0.9 --- somme des scores hub des pages citant 0 ou 2, qui sont : 9,10,11,12,13

Pour la page 3

$r_3^+ = 1-(1 - x)*(1 - y) / (1 - z) = 0.2323$

puisque

x=0.9 --- somme des scores hub des pages citant 0, qui sont : 9,10,11,12,13
y=0.2323 --- somme des scores hub des pages citant 3, qui sont : 10
z=0.9 --- somme des scores hub des pages citant 0 ou 3, qui sont : 9,10,11,12,13

Pour la page 4

$r_4^+ = 1-(1 - x)*(1 - y) / (1 - z) = 0.1742$

puisque

x=0.9 --- somme des scores hub des pages citant 0, qui sont : 9,10,11,12,13
y=0.1742 --- somme des scores hub des pages citant 4, qui sont : 9
z=0.9 --- somme des scores hub des pages citant 0 ou 4, qui sont : 9,10,11,12,13

Enfin, pour la page 0

$r_4^+ = 1-(1 - x)*(1 - x) / (1 - x) = 0.9$

puisque

x=0.9 --- somme des scores hub des pages citant 0, qui sont : 9,10,11,12,13

Ainsi,

r_0=0.5166
r_1 = 0.25
r_3= 0.1333
r_4=0.1
r_2=0

La variation globale du score de pertinence a été de 0.1 et on va estimer que le résultat est atteint.

On voit que, avantageusement, la deuxième itération a permis de remonter le score de la page 3 qui est citée par une page ayant un score hub (0,2323) bien meilleure que (0,1742) celui de la page qui cite la page 4.

Extension R^+

Initialement, pour ne traiter qu'un nombre réduit de pages, les scores de pertinence peuvent être calculés sur la base de R^- (si on avait pris $H=R^-$). Ceci ne sera alors qu'une approximation. En effet, pour que les scores soient corrects, il faudrait les calculer en se basant plutôt sur $H=R^{+-}$ ($R^{+-} = \bigcup_{P_i \in R^{-+}} B(P_i)$). Mais comme la constitution de R^{+-} est relativement coûteuse, on ne prendra qu'un sous-ensemble : on prendra pour R^{+-} seulement les pages pointant sur les pages de A qui ont un bon score.

Ainsi[39], on va ajouter une sous-étape avant la fin de l'étape 2.a :

2.a.1. Dans le cas où le score r_i^+ de la page courante (P_i de A) est suffisant[40], on recalcule r_i^+ après avoir inséré dans H les nouvelles pages de $B(P_i)$

$H \mapsto B(P_i) \bigcup H$.

[39] Plusieurs méthodes peuvent être utilisées ; nous présentons ici la préférée.

[40] (c'est-à-dire supérieur à un seuil choisi ; ce seuil pourra être fonction de la cardinalité courante de H, en effet plus on se rapproche de R^{+-} (e.g. H_{final}) plus le score calculé a des chances d'être déjà correct)

Défavoriser les pages moins fiables

On introduit un score autorité pour les pages de A et l'équation r_i^+ est $r = w_{i\cup K} \cdot a_i$ (plutôt que $r = w_{i\cup K}$). Le nouveau coefficient a_i permettra d'affaiblir les pages peu fiables (par le fait qu'ils sont peu populaires). En outre, l'équation sera plus cohérente dans la mesure où le score pertinence ne sera plus le même pour toutes les pages de la requête.

La procédure est maintenant la suivante :

1. Ce point est le même que celui du « Nouveau procédé de calcul de score de pertinence par l'amont » présenté plus haut.

2. (Re-)calculer les scores autorité :
   a. Pour chaque page P_i de A, en commençant par celles de K, associer un nombre a_i, son score autorité, égal à $\sum_j l_{ji} \cdot h_j$, où $l_{ji} = \begin{cases} 0 \text{ s'il n'y a pas de lien entre } P_j \text{ et } P_i \\ 1 \text{ s'il y a un lien entre } P_j \text{ et } P_i \end{cases}$.
   b. Si, pour certaines pages, a_i est suffisamment proche de sa valeur calculée précédemment (le cas échéant), et que les scores autorité des pages de K n'ont pas variés non plus, nous pouvons garder les anciennes valeurs a_i et r_i pour cette page, pour économiser les calculs.

3. (Re-)calculer les scores de pertinence :
   a. Pour chaque page P_i de A calculer r_i^+, égal à $w_{i\cup K} \cdot a_i$ et dans le cas où le résultat est négatif (cas d'une page antagoniste à R) neutraliser les liens entrants de manière à avoir $r_i^+ = 0$.

   L'homogénéité par l'amont w_S d'un ensemble S est défini comme suit:

   $$\overline{w_S} = \prod_{P \subset S} \overline{a_P}^{\sigma_P}, \text{ où}$$

   - $\sigma_P = \begin{cases} -1 \text{ si P contient un nombre pair de pages} \\ +1 \text{ sinon} \end{cases}$
   - $a_P = \sum_j h_j l_{jP}$ où

     $$l_{jP} = \begin{cases} +1 \text{ si } \exists P_i \in P \mid l_{ji} = +1 \\ 0 \text{ sinon} \end{cases},$$

     avec $l_{ji} = \begin{cases} 0 \text{ s'il n'y a pas de lien entre } P_j \text{ et } P_i \\ 1 \text{ s'il y a un lien entre } P_j \text{ et } P_i \end{cases}$

     i.e. l_{jP} est égal à 1 s'il y a un lien
     - d'une page P_j (de H)
     - à au moins une page P_i de P

     et zéro sinon.

     Ceci signifie tout simplement que a_P est le total des scores hub des pages (de H) qui pointent sur au moins une page de P (P est le sous-ensemble courant de S qui est considéré).

4. Poursuivre à partir du point 2.b du « Nouveau procédé de calcul de score de pertinence par l'amont » présenté précédemment.

Filtrer le Népotisme

Au point 3. ci-dessus, on remplace la définition de $\sum_j hl_{jP}$ par $\sum_j h\ell_{jP}$

$$\text{où } \ell_{jP} = \begin{cases} \min\left[1; \underset{i \in P}{\text{Max}} \overline{\overline{x_{ji} \cdot h_j}}\right] \text{ si } \exists P_i \in P \mid l_{ji} = +1 \\ 0 \text{ sinon} \end{cases}$$

(Voir les explications déjà données précédemment pour la méthode de filtrage du népotisme et pour la prise en compte du score des pages citantes dans cette méthode).

Puisqu'à chaque fois on doit calculer la proximité entre deux pages seulement, on préfère utiliser l'équation suivante (aussi déjà introduite):

$$x_{ij} = \frac{\left|B(P_i) \cap B(P_j)\right|}{\left|B(P_i) \cup B(P_j)\right|} \cdot \frac{\min\left[\left|B(P_i)\right|;\left|B(P_j)\right|\right]}{\max\left[\left|B(P_i)\right|;\left|B(P_j)\right|\right]}$$

On calcule le score de pertinence par l'aval comme suit:

Cadre général pour le traitement par l'aval

Nous supposons que R est « homogène par l'aval » – nous décrirons plus loin ce terme (voir ci-dessous la section « Décomposer une requête en parties homogènes »).

Exécuter le même algorithme, ainsi que son extension et ses variantes, en remplaçant R^- par R^+, R^{-+} par R^{+-}, $B(P_i)$ par $F(P_i)$ et $F(P_i)$ par $B(P_i)$.

Décomposer une requête en sous-requêtes homogènes

Rappelons que l'utilisateur fournit au système un ensemble R de pages auxquelles il est intéressé et éventuellement un ensemble de pages R_X de pages qu'il ne veut explicitement pas ($R \cap R_X = \varnothing$). Le système va identifier au sein de R au moins un groupe de pages « homogène » et va lancer une sous-requête séparée sur ce ou chaque groupe. Ces groupes sont appellés « kernel » (ou noyau).

Procédure :

1. Pour chaque page P_i de R, trouver $B(P_i)$, l'ensemble de pages citant P_i.
2. Trouver $R^- = \bigcup_{P_i \in R} B(P_i)$, l'ensemble de pages citant au moins une page de R.
3. Dans les pages de R qui ne sont pas encore dans un noyau (au début aucune ne l'est), trouver la page P_B ayant le plus large ensemble $B(P_B)$ de liens entrants[41], et créer un noyau contenant seulement cette page. Ce noyau est maintenant K_C, le noyau courant en construction (à tout instant il n'y en a qu'un seul).
4. Trouver les pages pertinentes par rapport à K_C (en utilisant le procédé de calcul de pertinence par l'amont) avec
   - $H=R^-$
   - $A=R$
   - $K=K_C$
   - $T= R_X$
5. Soit P_N la page de R, pas encore dans K_C, qui a le score de pertinence le plus élevé. Si son score de pertinence est inférieur à un score minimal fixé, créer un nouveau noyau contenant P_N (le noyau courant est maintenant complet). Sinon l'insérer dans K_C. Cette page est maintenant traitée. S'il reste encore au moins une page non encore traitée, retourner au point 3. Sinon continuer au point 6.
6. On a maintenant un ensemble de noyaux (sous-requêtes homogènes par l'amont) prêtes à être utilisées comme décrit dans ce document.

Bien entendu, la présente invention n'est nullement limitée aux formes de réalisation décrites ci-dessus et représentées sur les dessins, et l'homme du métier saura y apporter de nombreuses variantes ou modifications.

[41] Dans le cas où on a les scores autorité des pages, ou autre score de popularité, on préfère se baser plutôt sur eux.

REVENDICATIONS

1. Procédé pour déterminer des pages additionnelles pertinentes par rapport à un ensemble donné de pages de départ, caractérisé en ce qu'il comprend les étapes suivantes

a) identifier un ensemble de pages citantes constituées par toutes les pages ayant un lien vers au moins l'une des pages de départ,

b) former un ensemble de pages candidates constitué par l'ensemble des pages citées par les pages citantes,

c) pour chaque page candidate, calculer un score de pertinence de page candidate entre ladite page candidate et l'ensemble de pages de départ sur la base de l'existence de liens situés dans les pages citantes et dirigés vers la page candidate et vers les pages de départ, et sur la base également de scores de pertinence de pages citantes affectés à chacune des pages citantes,

d) pour chaque page citante, recalculer un score de pertinence de page citante sur la base de l'existence, dans la page citante en question, de liens vers les pages candidates et sur la base également des scores de pertinence de page candidate attribuées aux pages candidates à l'étape c),

e) répéter le cas échéant l'étape c) et le cas échéant une ou plusieurs fois l'étape d) puis l'étape c)

f) déterminer lesdites pages additionnelles pertinentes comme étant les pages candidates qui présentent les meilleurs scores de pertinence de page candidate.

2. Procédé selon la revendication 1, caractérisé en ce que le calcul de score de pertinence effectué à l'étape c) comprend le calcul d'une pluralité de sommes de scores de pertinence de pages citantes, chaque somme comprenant uniquement les scores de pertinences des pages citantes comprenant un lien vers une page donnée constituée par la page candidate ou une page de départ.

3. Procédé selon la revendication 2, caractérisé en ce qu'il comprend également le calcul d'au moins une somme de scores de pertinence de pages citantes, chaque somme comprenant uniquement les scores de pertinences des pages citantes comprenant un lien vers l'une parmi un ensemble d'au moins deux pages données, cet ensemble comprenant la page candidate et au moins une page de départ.

4. Procédé pour déterminer des pages additionnelles pertinentes par rapport à un ensemble donné de pages de départ, caractérisé en ce qu'il comprend les étapes suivantes

a) identifier un ensemble de pages citées constituées par toutes les pages ayant un lien depuis au moins l'une des pages de départ,

b) former un ensemble de pages candidates constitué par l'ensemble des pages citant les pages citées,

c) pour chaque page candidate, calculer un score de pertinence de page candidate entre ladite page candidate et l'ensemble de pages de départ sur la base de l'existence de liens situés dans la page candidate et dans les pages de départ et dirigés vers les pages citées, et sur la base également de scores de pertinence de pages citées affectés à chacune des pages citées,

d) pour chaque page citée, recalculer un score de pertinence de page citée sur la base de l'existence, dans la page citée en question, de liens depuis les pages candidates et sur la base également des scores de pertinence de page candidate attribuées aux pages candidates à l'étape c),

e) répéter le cas échéant l'étape c) et le cas échéant une ou plusieurs fois l'étape d) puis l'étape c)

f) déterminer lesdites pages additionnelles pertinentes comme étant les pages candidates qui présentent les meilleurs scores de pertinence de page candidate.

5. Système de navigation parmi des ressources d'information, chaque ressource comprenant au moins un lien activable dans un premier mode par un dispositif d'entrée pour provoquer l'accès à une autre ressource d'informations désignée par un identificateur de ressource associé à ce lien, caractérisé en ce qu'au moins certaines ressources comprennent au moins un lien activable dans un second mode à l'aide d'un dispositif d'entrée pour envoyer à un moteur de recherche de nouvelles ressources d'informations une requête de recherche contenant l'identificateur de ressource associé au lien en question.

6. Système selon la revendication 5, caractérisé en ce que le dispositif d'entrée est apte à activer le lien simultanément dans les premier et second modes.

7. Système selon la revendication 5, caractérisé en ce que l'activation du lien dans le second mode est apte à provoquer l'affichage d'une requête pré-existante, à laquelle l'identificateur de ressource associé au lien en question est susceptible d'être ajouté.

8. Système selon les revendications 6 et 7 prises en combinaison, caractérisé en ce que l'activation du lien dans le second mode est apte à afficher, en plus de la requête pré-existante, la ressource d'informations désignée par ledit identificateur de ressource.

9. Système de recherche de nouvelles ressources d'information à partir de ressources d'informations existantes, caractérisé en ce qu'il comprend un moteur de recherche basé sur l'analyse de liens entre différentes ressources et acceptant en entrée une requête comprenant une série d'identificateurs de ressources, caractérisé en ce qu'il comprend un moyen de sélection d'identificateurs apte à mémoriser un ensemble d'identificateurs (URI) de ressources sélectionnés les uns après les autres par un utilisateur, et un moyen générateur de requête activable par l'utilisateur pour engendrer une requête contenant l'ensemble des identificateurs précédemment sélectionnés à destination du moteur de recherche.

10. Système selon la revendication 6, caractérisé en ce que le moyen de sélection est apte à mémoriser les identificateurs sélectionnés de manière rémanente, de telle sorte que le moyen de sélection puisse être mis en œuvre de façon espacée dans le temps en vue de la génération d'une même requête.

11. Procédé de recherche de nouvelles ressources d'information à partir de ressources d'informations existantes, caractérisé en ce qu'il comprend la mise en œuvre d'un moteur de recherche basé sur l'analyse de liens entre différentes ressources et acceptant en entrée une requête comprenant une série d'identificateurs de ressources et en ce qu'il comprend les étapes suivantes :

- sélection d'identificateurs (URI) de ressources les uns après les autres par un utilisateur ;
- génération d'une requête contenant l'ensemble des identificateurs précédemment sélectionnés à destination du moteur de recherche.

12. Procédé de recherche de nouvelles ressources d'information à partir de ressources d'informations existantes, caractérisé en ce qu'il comprend la mise en œuvre d'un moteur de recherche basé sur l'analyse de liens entre différentes ressources et acceptant en entrée une

requête comprenant une série d'identificateurs de ressources et en ce qu'il comprend les étapes suivantes :

- génération d'une requête contenant un ensemble d'identificateurs de ressources précédemment mémorisés dans un même groupe d'identificateurs de ressources propre à un utilisateur, à destination du moteur de recherche,
- génération d'une signalisation à l'attention de l'utilisateur lorsqu'au moins un nouvel identificateur de ressource appartenant au groupe en question a été trouvé par le moteur.

13. Procédé selon la revendication 12, caractérisé en ce que chaque groupe d'identificateurs de ressources est représenté par un objet graphique sur un dispositif d'affichage de l'utilisateur, et en ce que ladite signalisation est réalisée au moins par changement d'apparence de cet objet graphique.

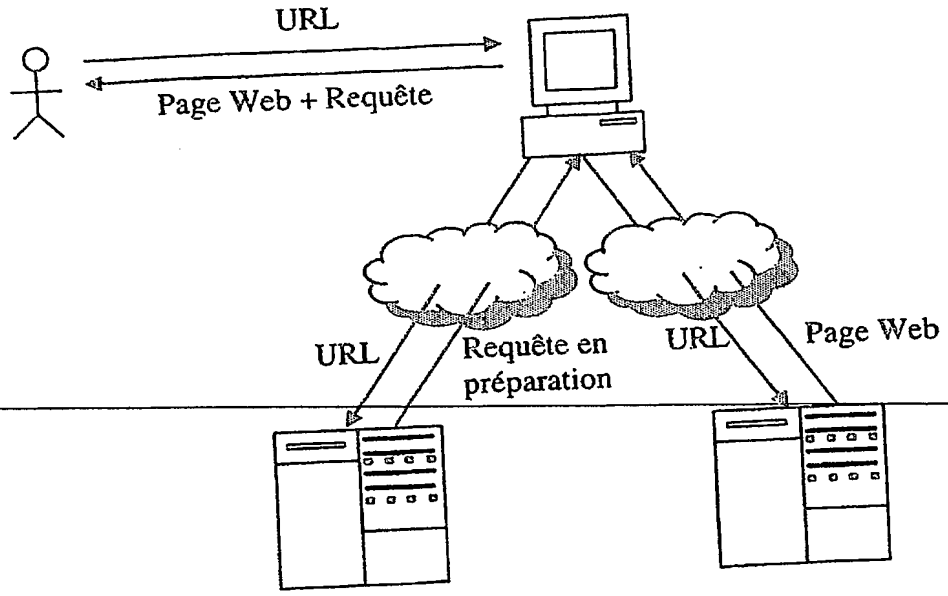
URL
Page Web + Requête
URL
Requête en
préparation
URL
Page Web

Fig. 1a

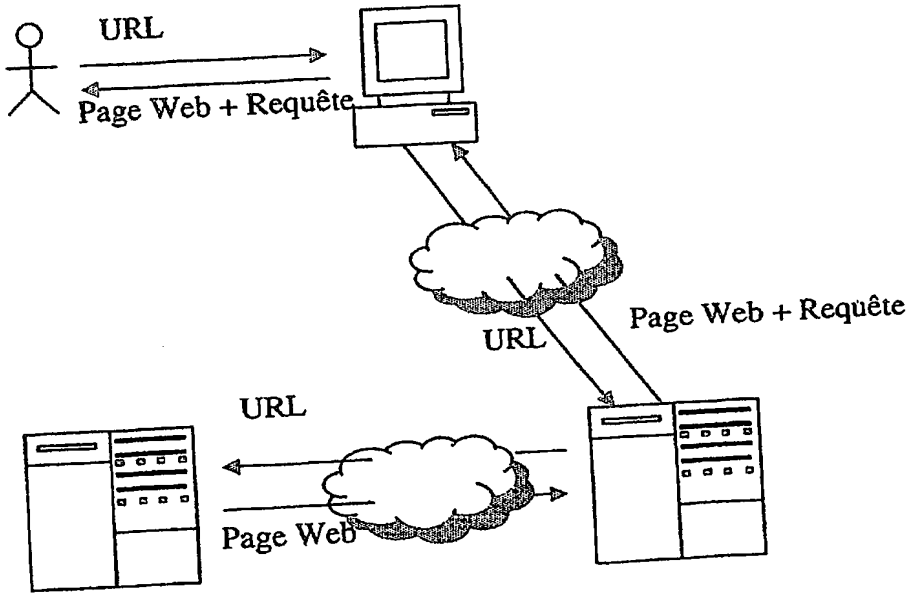
URL
Page Web + Requête
Page Web + Requête
URL
URL
Page Web

Fig. 1b

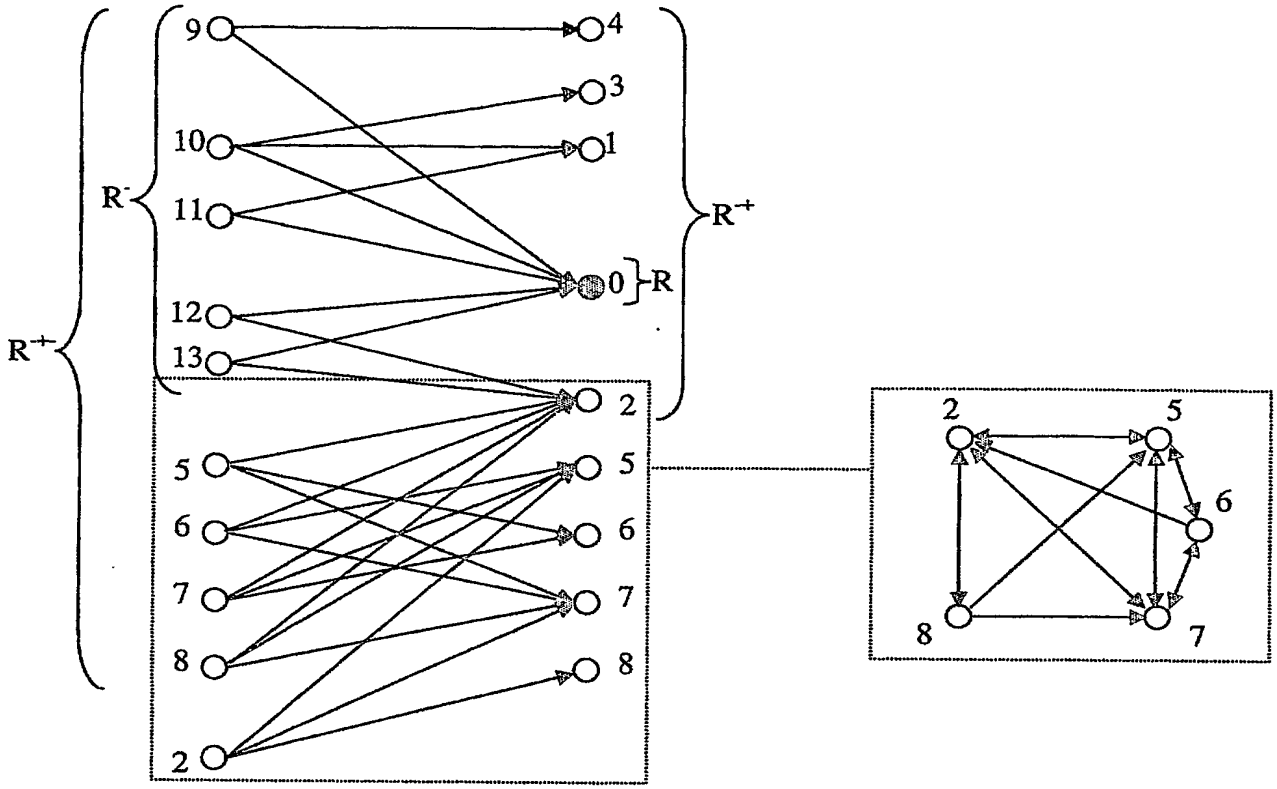
9
10
11
12
13
5
6
7
8
2
4
3
1
0
2
5
6
7
8
R
R⁻
R⁺
R⁺
2
5
6
8
7

Fig. 2

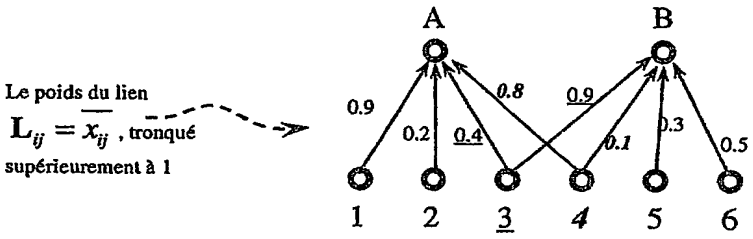
Le poids du lien
$L_{ij} = \overline{x_{ij}}$, tronqué
supérieurement à 1
A
B
0.9
0.2
0.4
0.8
0.9
0.1
0.3
0.5
1
2
3
4
5
6

Fig. 3

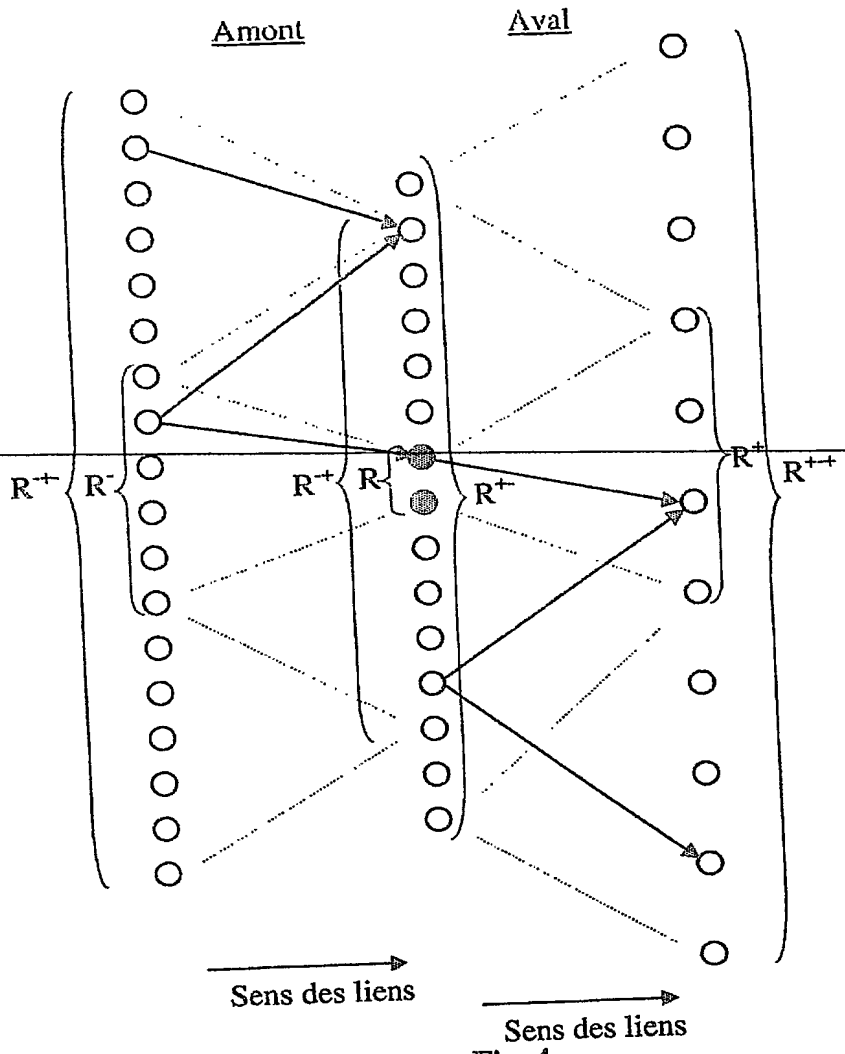
Amont
Aval
R-+
R-
R+
R
R+-
R+
R+-
Sens des liens
Sens des liens

Fig. 4

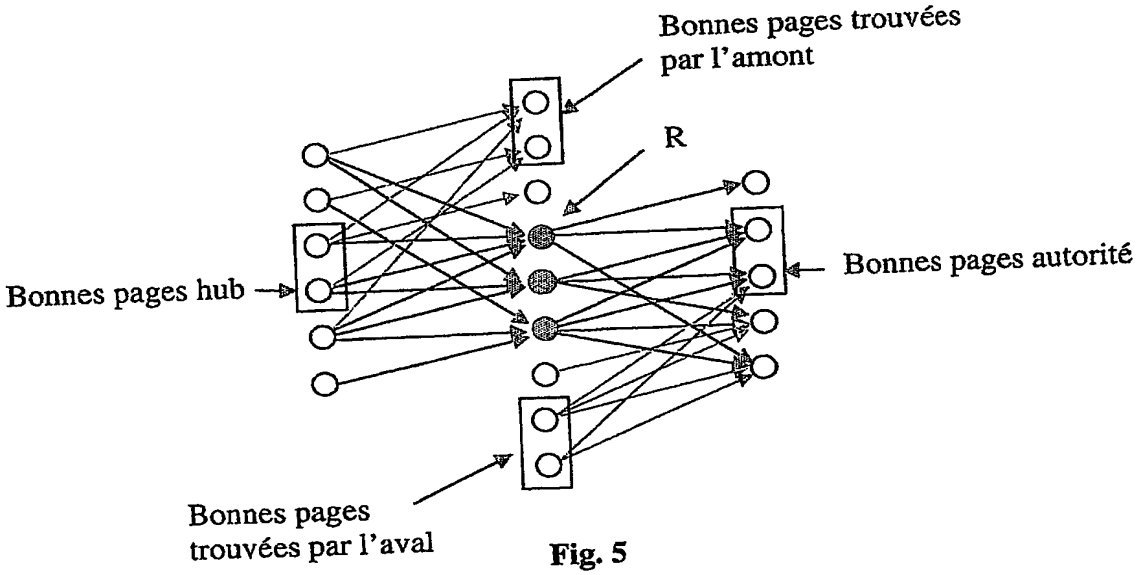
Bonnes pages trouvées par l'amont
R
Bonnes pages hub
Bonnes pages autorité
Bonnes pages trouvées par l'aval

Fig. 5